

---

# Evaluating Crowdsourcing Participants in the Absence of Ground-Truth

---

**Ramanathan Subramanian**  
Northeastern Univ.  
Boston MA 02215

**Rómer Rosales**  
LinkedIn  
Mountain View CA 94043

**Glenn Fung**  
Siemens  
Malvern PA 19355

**Jennifer Dy**  
Northeastern Univ.  
Boston MA 02215

## Abstract

Given a supervised/semi-supervised learning scenario where multiple annotators are available, we consider the problem of identification of adversarial or unreliable annotators.

## 1 Introduction

Data can be acquired, shared, and processed by an increasingly larger number of entities, in particular people. The distributed nature of this phenomenon has contributed to the development of many crowdsourcing projects. This scenario is prevalent in most forms of expert/non-expert group opinion and rating tasks (including many forms of internet or on-line user behavior), where a key element is the aggregation of observations-opinions from multiple sources.

This data acquisition setting has created a number of interesting problems and opportunities for machine learning and data modeling among other disciplines. Efficiently aggregating and obtaining knowledge from multiple distributed sources is far from being a solved problem. From a machine learning point of view, we treat observations and opinions as data points and participant label/annotations respectively. In this sense, multiple participants (annotators) may assign conflicting labels to the same data point. Clearly, participant expertise varies and some data points may be considered *easier* for certain participants relative to others. Thus, the selection of and trust placed on participants plays a crucial role on crowdsourcing outcomes.

This paper describes an approach for learning to label/annotate by modeling participant expertise, and focuses on a novel approach to identify unhelpful or adversarial participants. The paper presents an approach for participant (also referred to as annotators) evaluation based on a multi-annotator statistical model and provides experimental validation of the approach with respect to number of adversaries, degree of adversarial behavior, and effect on the classification accuracy.

## 2 Problem Setup

Let the available data collected from multiple annotators be given by a set of points  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  drawn independently from an input distribution. Some of these data points may have been labeled by one or more annotators. We denote by  $y_i^{(t)}$  the label for the  $i$ -th data point given by annotator  $t$  and let  $Y = \{y_i^{(t)}\}_{it}$ . We let  $\mathbf{x}_i$  and  $y_i^{(t)}$  be random variables in an input and label space respectively and introduce additional random variables  $Z = \{z_1, \dots, z_N\}$  to represent the *true* but not observed label for the data points.

Annotators may label certain data points more reliably than others because of the properties of the data point itself. For example, a noisy data point maybe more difficult to label by all annotators, but also a data point

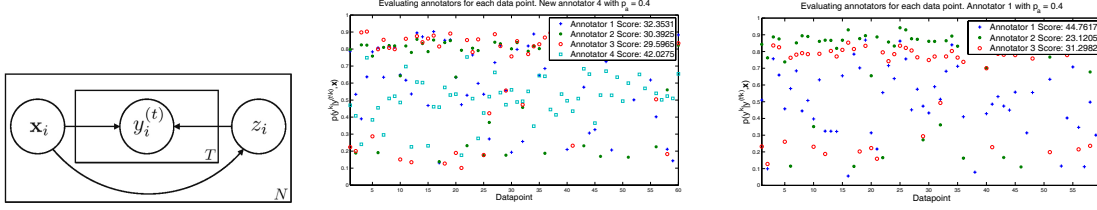


Figure 1: Left: Graphical Model for  $x$ ,  $y$ , and  $z$  respectively inputs, annotator-specific labels, and ground truth label (left). Center-Right: Probability corresponding to Eq.1 for each data point: (center) one adversary introduced with  $p_a = 0.4$ , (right) annotator\_1's labels randomly flipped with probability  $p_a = 0.4$ .

may be more familiar to certain annotators and thus the annotator-specific accuracy would vary. Thus, we let the annotation  $y^{(t)}$ , provided by labeler  $t$ , depend on the true unknown label  $z$  but also on the input  $x$ .

This is in contrast with the related work in [1,2] where annotators are assumed uniformly accurate/inaccurate irrespective of the data point being labeled. We instead follow the model introduced in [3] where the dependency  $x \rightarrow y^{(t)} \leftarrow z$  is explicit. This can be summarized by the graphical model in Fig. 1-Left.

We focus on binary classification and use a Bernoulli distribution to model the ability of the annotators to provide the correct label:  $p(y_i^{(t)} | x_i, z_i) = (1 - \eta_t(x))^{y_i^{(t)} - z_i} \eta_t(x)^{1 - |y_i^{(t)} - z_i|}$ , where the probability of success  $\eta_t(x)$  is a function of  $x$ , thus allowing the annotator accuracy to depend on the input data point coordinates. We let the true label depend on  $x$  via a logistic regression model for  $p(z_i = 1 | x_i)$ .

### 3 Evaluating Annotators

Consider any single data point. If we knew the ground-truth, then we could straightforwardly evaluate the annotator accuracy. What if we do not have access to the ground-truth (it does not exist or is expensive to obtain)? The following proposed distribution provides a way to evaluate an annotator without reliance on ground-truth. The basic idea is to evaluate a conditional distribution measuring how predictable an annotator label is conditioned on other annotators:

$$p(y^{(k)} | \{y^{(t \setminus k)}\}, x) = \frac{p(\{y^{(t)}\} | x)}{p(\{y^{(t \setminus k)}\} | x)} = \frac{\sum_z p(\{y^{(t)}\} | z, x) p(z | x)}{\sum_z p(\{y^{(t \setminus k)}\} | z, x) p(z | x)} \quad (1)$$

We note that if the ground-truth is given (along with the input data), the annotators are mutually independent and  $p(y^{(k)} | \{y^{(t \setminus k)}\}, x) = p(y^{(k)} | x)$ , as expected. A related way to understand Eq. 1 is as a measure of how much an annotator's label, for a given data-point, can be correctly inferred from that of other annotators.

### 4 Experiments

We start with two medical multi-annotator datasets: (1) A breast cancer dataset where mammograms have been labeled independently by three doctors. Ground-truth has been obtained through a biopsy, not available to the algorithm nor the participant doctors. (2) A concept inference dataset where non-experts read a passage of text, from de-identified electronic medical records, to assess whether the patient has atrial fibrillation, a type of heart arrhythmia. Ground-truth was given by expert assessment. The datasets contained 75 and 1057 datapoints respectively. For a more controlled experiment, for both datasets we additionally generated  $M$  helpful/adversarial annotators. The labels given by these were generated as the true labels but flipped with

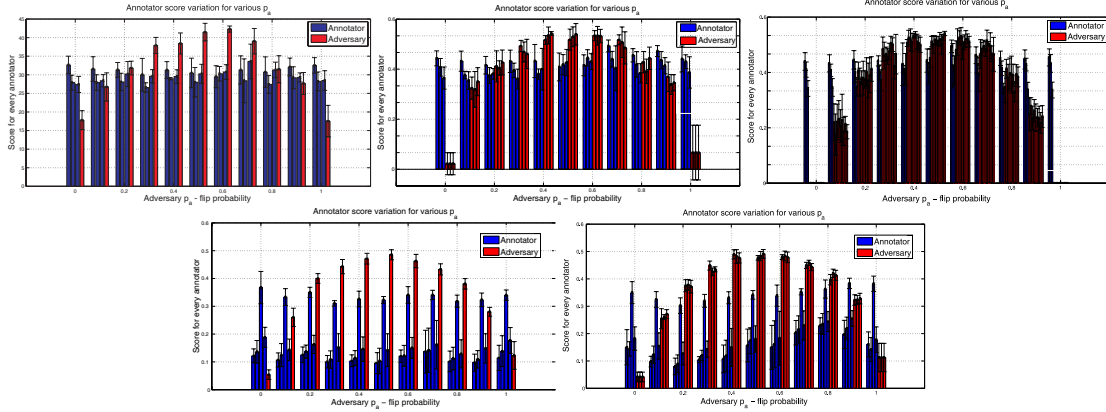


Figure 2: Scores for each annotator when 1 (left), (center), and 9 (right) adversaries are introduced for the Breast Cancer (top) and Atrial Fibrillation (bottom) datasets.

probability  $p_a$  (thus,  $p_a$  effectively determines the nature of the annotator). For all experiments we trained the multi-labeler model described in Sec. 2 to build the distributions required by Eq. 1.

In order to illustrate the properties of the approach, Fig. 1 shows breast cancer data points labeled by various annotators plotted against the value of  $p(y^{(k)}|\{y^{(t \setminus k)}\}, \mathbf{x})$  for some specific annotator  $k$  (Eq. 1). In Fig. 1-center we flipped the label of the first annotator with  $p = 0.4$  causing the conditional probability of the labels given by this annotator to decrease, as expected. In Fig. 1-right a fourth adversarial annotator with  $p_a = 0.4$  was introduced. This annotator obtains a lower conditional probability value comparatively, indicating that it is relatively more difficult to predict its label compared with the original three annotators.

We additionally measured the combined impact of the number of adversarial annotators and the strength of  $p_a$ . For this we defined an *adversarial score* for an annotator as the sum (over data points) of the negative log conditional probabilities defined by Eq. 1. This is comparable with the negative log-likelihood measure frequently utilized in learning tasks. Thus, if an annotator’s labels are difficult to infer based on the rest of the annotator labels, then the conditional probabilities above will be generally low. The lower these probabilities the larger the score; thus, a large score is indicative of an annotator that is less trustworthy.

Fig. 2 shows this score while varying  $p_a$  and the number of helpful/adversarial annotators. We noted various properties. The scores of the non-adversarial annotators remained fairly constant. This indicates that the scoring scheme is robust to the number of adversaries and the strength of  $p_a$  and could be used for detecting adversaries. The score curves for the adversaries are symmetric about  $p_a = 0.5$ , which is consistent with the notion that a completely random annotator corresponds to the worst case (independent) adversary since no information can be gained from its labels.

We additionally tested this approach with various UCI datasets. Since these datasets do not employ multiple annotators, all the non-adversarial annotators (blue) were generated at random based on the true label, while the adversarial annotators (red) were generated as before. Results, shown in Fig. 3 are consistent with the previous experiments and further confirm the benefits of the approach in a more controlled setting.

Finally, Figs. 4-5 show how the effect of adversarial annotators and the value of  $p_a$  can impact the performance (AUC) of the model. We can observe that the AUC decreases considerably only when a large proportion of the adversarial annotators consistently flip the correct label. This indicates that the model can remain accurate under a wide range of adversarial influence.

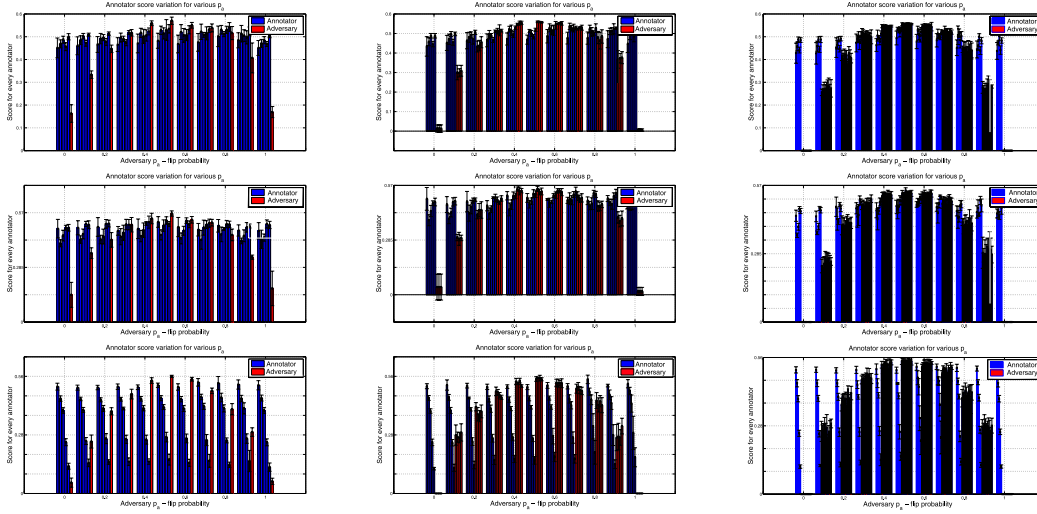


Figure 3: Scores for each annotator when 1 (left), 3 (center), and 9 (right) adversaries are introduced for the Housing (top), Ionosphere (middle), and Glass (bottom) datasets.

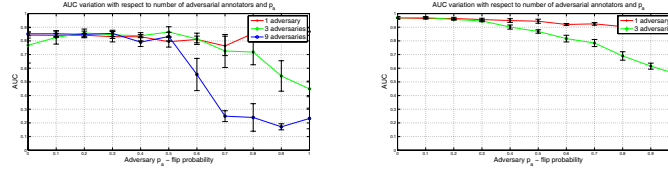


Figure 4: AUC for various  $p_a$  values and varying number of helpful/adversarial annotators for the Breast Cancer and Atrial Fibrillation datasets.

## References

- [1] V. C. Raykar et al. Learning from crowds. *J. Machine Learning Research*, 11(Apr):1297 - 1322, 2010
- [2] R. Snow et al. Evaluating non-expert annotations for natural language tasks. In Proc. *EM-NLP*, 2008.
- [3] Y. Yan et al. Modeling annotator expertise: Learning when everybody knows a bit of something. In Proc. *Int. Conference on Artificial Intelligence and Statistics*, 2010.

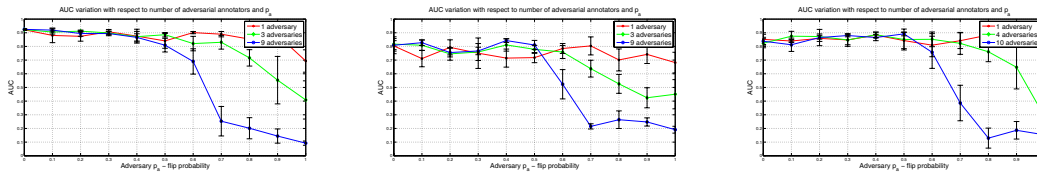


Figure 5: AUC for various  $p_a$  values and varying number of helpful/adversarial annotators for the Housing, Ionosphere, and Glass datasets.